DOCUMENT RESUME

ED 271 452 SP 027 779

TITLE	Teacher Performance Appraisal System Training: A Report of Outcomes.
INSTITUTION	North Carolina State Dept. of Public Instruction, Raleigh.
PUB DATE	Mar 86
NOTE	26p.; For a related document, see SP 027 780.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Evaluation Methods; *Program Effectiveness; State Standards; Teacher Effectiveness; *Teacher Evaluation; Teaching Skills; *Training Methods
IDENTIFIERS	South Carolina; *Teacher Performance Appraisal Instrument

ABSTRACT

The State Department of Public Instruction of North Carolina has implemented a training program designed to increase the skills of principals, observer/evaluators, and others in the application of the Teacher Performance Appraisal Instrument (TPAI). This report focuses on the effectiveness of that training. In all, a total of almost 900 individuals from 140 schools have been trained. The TPAI requires the evaluator to make judgments about the quality of a teacher's performance on five functions of the teaching act: (1) managing instructional time; (2) managing student behavior; (3) instructional presentation; (4) monitoring instruction; and (5) providing instructional feedback. This report explains how the rating system is used for these functions, and describes the activities of the 27 training workshops. An analysis is presented of how each of the functions should be observed. A brief discussion is given on differing abilities of individuals in utilizing the TPAI, and on the relative success of the training sessions. Data gathered from the workshops on participant performances are presented. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E Brumback

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TABLE OF CONTENTS

	Page
Background	1
Results	2
Analysis of Functions	7
Success of Individual Participants	15
Conclusion	19

List of Tables

Table 1:	Participants by Region and Job-Role	3
Table 2:	Participants Performance on Post-Test	6
Table 3:	Distribution of Participants' Choices by Function	7
Table 4:	Distribution of Responses on Function 1 by Region	8
Table 5:	Distribution of Responses on Function 2 by Region	9
Table 6:	Distribution of Responses on Function 3 by Region	10
Table 7:	Distribution of Responses on Function 4 by Region	11
Table 8:	Distribution of Responses on Function 5 by Region	12
Table 9:	Summary Results by Function	13
Table 10:	Distribution of Scores on Acceptable/Unacceptable Ratings (Percentage of Participants Describing Performance as Acceptable/Unacceptable)	14
Table 11:	Number and Percentage of Participants Ratings All 5 Functions Within Acceptable Range and Outside Acceptable Range, by Region and By Total	15
Table 12:	Number/Percentage of Participants in Job Roles Within Acceptable/Unacceptable Ranges by Regions	17
Table 13:	Number of Participants Out on 1 and Out on 2 or More	18

Background

The Personnel Services Area of the State Department of Public Instruction is undertaking a number of coordinated efforts that all share a common goal: the enhancement of teaching that will lead to educational improvement for boys and girls throughout North Carolina. While this goal is shared by educators working at all levels of organization, by citizens, and by political leaders, in the Personnel Services Area, the goal is made concrete largely through the development and implementation of performance appraisal measures that emphasize specific teaching practices that have been associated in research literature with increased student achievement.

A number of activities have been undertaken by staff of the Personnel Services Area to achieve this goal. First, a training program has been designed that increases teachers' knowledge and awareness of 28 specific practices associated with increased achievement. This Effective Teaching Training has been provided by State Department of Public Instruction staff to hundreds of individuals throughout the state. Similarly, training that was designed to increase the skills of principals, observer-evaluators, and others in the application of the observation and evaluation cycle has been implemented. This report will focus on the effectiveness of that training.

During November and December 1985, 25 training sessions on the Teacher Performance Appraisal System were held throughout North Carolina. Representatives of every local education agency in the state were allocated opportunities to participate in this training. In January 1986, State Department of Public Instruction conducted two additional training sessions that ensured that local education agencies participating in the pilot Career Ladder Development program had sufficient opportunities to ensure training for appropriate administrators and teachers. In all, a total of almost 900 individuals from 140 school units were trained.

Although individual school units were permitted to sponsor whomever they chose to attend the training, State Department of Public Instruction suggested that the training was especially appropriate for principals and observer-evaluators in those districts participating in the Career Ladder pilot program. Indeed, the legislation authorizing the Career Ladder pilot program specifies that teachers are to be observed and evaluated by trained administrators. Therefore, State Department of Public Instruction allocated more spaces to personnel from pilot Local Education Agencies than to persons from non-participating Local Education Agencies. However, because State Department of Public Instruction staff will not be able to provide training in use of the new observation-evaluation system to every principal in every school system, a turnkey training model was adopted in the State Department of Public Instruction-sponsored workshops. That is, participants trained by State Department of Public Instruction are expected to provide training through local workshops to those of their colleagues who did not receive training directly from State Department of Public Instruction. Each district was also provided master sets of all training materials for local reproduction.

Results

Table 1 shows the number of individuals from each educational region who participated in the 27 State Department of Public Instruction-sponsored workshops. Participants are identified in terms of region and job-role: principal, observer-evaluator, other (a term that encompasses superintendents, central office personnel, assistant principals, and teachers). In addition, a number of individual participants were not identified by job role; these are aggregated in the column labelled "Unidentified". The last column in Table 1 indicates the total number of persons from each region who received training

and the percentage of that number of the total 896 persons trained in the 27 workshops. Similarly, the last number in each column represents the aggregation of all role participants, with the percentage of the total participants.

TABLE 1

Participants By Region and Job-Role

Role / Region	Principal	Observer/Evaluator	Other	Unidentified	Total (%)
1	60 (57%)	1 (1%)	43 (41%)	2 (1%)	106 (12%)
2	31 (31%)	17 (17%)	50 (51%)	1 (1%)	99 (11%)
3	53 (50%)	4 (4%)	44 (42%)	4 (4%)	105 (12%)
4	56 (52%)	5 (5%)	45 (42%)	1 (1%)	107 (12%)
5	46 (48%)	7 (7%)	40 (42%)	3 (3%)	96 (11%)
6	45 (51%)	0 (0%)	42 (47%)	2 (2%)	89 (10%)
7	97 (64%)	13 (9%)	37 (24%)	5 (3%)	152 (17%)
8	53 (56%)	14 (15%)	18 (19%)	9 (10%)	94 (10%)
Makeup	11 (23%)	1 (2%)	35 (73%)	1 (2%)	48 (5%)
Total	452 (50%)	62 (7%)	354 (40%)	28 (3%)	896 (100%)

Slightly more than 50 percent of all persons trained identified themselves as school principals. Interestingly, in Regions 1 and 6, and during the makeup sessions, fewer than 2 percent of the participants identified themselves as observer-evaluators. Since observer-evaluators are required to receive this training, we suspect that individuals in this category did not identify themselves as such.

In designing the training, State Department of Public Instruction staff felt that each training session could reasonably accommodate between 30 and 36 persons. In fact, the session average was about 33, a number that permitted maximum participation with maximum efficiency of training.

In addition to collecting quantity data, State Department of Public Instruction was able to collect data that report quality of training. Quality can be defined in at least two ways: (1) quality of the training experience

and (2) quality of training outcome. This report will confine itself to analysis of quality of the second dimension. The primary goal of training was the development of skill as an observer and evaluator of teaching practices. To determine success, State Department of Public Instruction analyzed the results of completion of evaluations of a video-taped training episode.

The Teacher Performance Appraisal Instrument requires the evaluator to make judgments about the quality of a teacher's performance on five functions of the teaching act:

1) Managing Instructional Time
2) Managing Student Behavior
3) Instructional Presentation
4) Monitoring Instruction
5) Providing Instructional Feedback

The evaluator, using data collected from observation of classroom performance, assigns a rating of 1-6 for each of these functions. Thus, a five-number score for performance is created as shown in this example:

Function	1	2	3	4	5
Score (Choose 1-6)	3	3	3	3	3

In the example above, the teacher's performance represents "at standard" teaching, since each function was rated 3. Because social science is not as exact as natural science, our evaluation system has a built-in tolerance, sometimes referred to as the standard error of measure. Because our scale moves in whole numbers, a tolerance of ±1 is deemed acceptable. That is, any evaluator can award a score of ± on any function and still be in the acceptable

range. Put another way, these ratings by three evaluators of the same teacher's performance are all equally acceptable:

Function	1	2	3	4	5
Score	2	2	2	2	2

Function	1	2	3	4	5
Score	3	3	3	3	3

Function	1	2	3	4	5
Score	4	4	4	4	4

These ratings are functionally equivalent. We will return to this point in our discussion below. For now, however, it is enough to understand that this tolerance is built into the system.

In training, a video-tape of a ninth-grade English lesson was used as a post-test. After four days of training, participants were shown the tape and asked to record their observations and evaluate the teaching. The performance was normed by qualified State Department of Public Instruction staff who had observed the tape and who had come to agreement on the scores. Table 2 below shows the correct or normed score for each function as demonstrated on the video-tape test. It also shows the mode for participant responses. The mode is a statistic representing the most frequently chosen answer. Two other facts are shown in Table 2: the number and percentage of participants choosing the correct answer and the number and percentage of participants choosing an answer in the acceptable range (correct score ±1).

TABLE 2

Participants Performance on Post-Test

Function	Correct Score	Mode	# Correct	(%)	# Acceptable	(%)
1	2	3	320	(36%)	775	(88%)
2	2	2	633	(72%)	863	(98%)
3	2	3	308	(35%)	749	(85%)
4	3	2	322	(36%)	850	(96%)
5	3	3	527	(60%)	844*	(96%)

(*Number does not equal 896 because of individual failure to respond to each item.)

On every function, more than a third of all participants chose the correct response, and 85 percent or more were within the acceptable range. On Functions 1 (time management), 3 (instructional presentation), and 4 (instructional monitoring), however, the mode was different from the norm score. This indicates a need to exercise caution when accepting the measure of success indicated by the high percentage of people choosing scores within the acceptable range.

A large amount of training was devoted to helping participants learn to value teaching practices in a similar way. That is, once each participant learned to recognize evidences of the teaching practices on which the evaluation rests, it was important that participants place a similar value on the demonstration of those practices by the teacher observed. Table 3 shows the distribution of participant choices for each teaching function.

TABLE 3

Distribution of Participants' Choices by Function

Score / Function	1	2	3	4	5	6	N
1	11 (1%)	320 (36%)	444 (50%)	95 (11%)	14 (2%)	1 (1%)	885
2	77 (8%)	633 (72%)	153 (17%)	20 (2%)	2 (1%)	0 (0%)	885
3	31 (4%)	308 (35%)	410 (46%)	113 (13%)	19 (2%)	1 (1%)	882
4	33 (4%)	488 (55%)	322 (36%)	39 (4%)	3 (1%)	1 (1%)	886
5	14 (2%)	164 (19%)	527 (60%)	153 (17%)	23 (3%)	1 (1%)	882

These data present a picture of participant performance on each function. As we have seen, most participants were within acceptable limits on each function. However, within any single function, there was a distribution of responses across at least five of the six quality points. Moreover, it would appear that one function evoked more nearly unanimous choices than did any of the other functions. An examination in more detail of each function will help us understand better participants' performance in the training.

Analysis of Functions

Function 1 is concerned with the teacher's management of instructional time. The observer notes whether the teacher begins class promptly, whether materials for learning are ready, and how the teacher gets--and keeps--students on task throughout the lesson. These practices, of course, interact with practices in other functions and it is this interaction that often causes the observer difficulty in recording specific instances of the practice. Table 4 below shows the number and percentage of participants, by region, who chose any one of the six quality points on Function 1.

TABLE 4

Distribution of Responses on Function 1 by Region

Score / Region	1 N (%)	*2 N (%)	3 N (%)	4 N (%)	5 N (%)	6 N (%)	Total
1	1 (1%)	37 (36%)	46 (45%)	16 (16%)	3 (3%)	0 (0%)	103
2	1 (1%)	23 (24%)	57 (59%)	14 (14%)	1 (1%)	1 (1%)	97
3	2 (2%)	56 (53%)	40 (38%)	5 (5%)	3 (3%)	0 (0%)	106
4	2 (2%)	34 (32%)	49 (46%)	18 (17%)	3 (3%)	0 (0%)	106
5	1 (1%)	32 (34%)	48 (51%)	11 (12%)	2 (2%)	0 (0%)	94
6	3 (3%)	31 (35%)	50 (56%)	5 (6%)	0 (0%)	0 (0%)	89
7	1 (1%)	75 (50%)	70 (46%)	5 (3%)	0 (0%)	0 (0%)	151
8	0 (0%)	19 (20%)	60 (65%)	12 (13%)	2 (2%)	0 (0%)	93
Makeup	0 (0%)	13 (28%)	24 (52%)	9 (20%)	0 (0%)	0 (0%)	46
Total	11 (1%)	320 (36%)	444 (50%)	95 (11%)	14 (2%)	1 (1%)	885

For Tables 4-8 the asterisk indicates the correct response, which was chosen by up to 53 percent of participants from any region. It would be interesting to know, however, why only 20 percent of participants in Region 8 chose this response. This question is more than academic if we are interested in ensuring equity across the state of North Carolina with respect to the use of the teacher appraisal instrument. We might begin to answer our question if we knew, for instance, that educators in Region 3, where 53% of participants chose the correct responses, had received more training in the principles of time management than had educators in Region 8. The answer to this question, of course, would not lead us to making a value judgment about Region 3 educators as compared with those of Region 8. It would, however, help us to understand the necessary level of training/skill needed to evaluate time management. The same point, as we shall see, can be made about each of the functions.

Function 2 is concerned with the teacher's ability to manage students' behavior. Strategies and practices related to discipline are subsumed in this function. Perhaps it is not surprising that participants were more in agreement on this function than on any other. First, we can logically construct the argument that administrators deal directly with significant behavior problems. They are, therefore, more apt to recognize teacher behavior that have the effect of increasing or decreasing these behavior problems. Second, there is anecdotal evidence to support the contention that a great deal of effort has recently been expended on training teachers and principals in strategies that increase the likelihood of acceptable student discipline. Programs like Assertive Discipline, Teacher Effectiveness Training, COET and TESA are some of these programs that have received attention from educators throughout the state. Each of them concerns itself, to some extent, with management of student behavior. Table 5 shows the distribution of responses on Function 2.

TABLE 5

Distribution of Responses on Function 2 by Region

Score / Region	1 N (%)	*2 N (%)	3 N (%)	4 N (%)	5 N (%)	6 N (%)	Total
1	14 (14%)	72 (70%)	14 (14%)	3 (3%)	0 (0%)	0 (0%)	103
2	5 (5%)	62 (64%)	25 (26%)	5 (5%)	0 (0%)	0 (0%)	97
3	12 (11%)	80 (75%)	11 (10%)	1 (1%)	2 (2%)	0 (0%)	106
4	7 (7%)	72 (68%)	24 (23%)	3 (3%)	0 (0%)	0 (0%)	106
5	9 (10%)	64 (68%)	18 (19%)	3 (3%)	0 (0%)	0 (0%)	94
6	10 (11%)	70 (79%)	9 (10%)	0 (0%)	0 (0%)	0 (0%)	89
7	13 (9%)	124 (83%)	13 (9%)	0 (0%)	0 (0%)	0 (0%)	150
8	5 (5%)	69 (73%)	17 (18%)	3 (3%)	0 (0%)	0 (0%)	94
Makeup	2 (4%)	20 (43%)	22 (48%)	2 (4%)	0 (0%)	0 (0%)	46
Total	77 (8%)	633 (72%)	153 (17%)	20 (2%)	2 (1%)	0 (0%)	885

As this table shows, not only did a greater percentage of participants select the correct answer but also the percentage of participants within the acceptable range was quite high. Interestingly, participants at workshops in Regions 1, 3, 6, and 7 who did not choose the correct response were almost perfectly divided between the two ±1 categories.

Function 3 relates to Instructional Presentation. There are a total of 11 practices in this function, some of which relate to research on lesson design (3.1: begins with a review; 3.2: states objective; and 3.11: brings closure to lesson) while others relate to lesson pace, teacher's use of language, use of examples and demonstrations, etc. If Function 3 related only to instructional design, we might hypothesize that those educators who have worked closely with proponents of Madeline Hunter's work in instructional presentation might be more successful on this aspect of the evaluation instrument. However, Function 3 is larger than just the "six-step lesson design" and examination of participants' responses does not support the hypothesis that some region's educators are more or less successful at identifying and valuing instructional presentation than are others. Table 6 presents the data related to performance on Function 3.

TABLE 6

Distribution of Responses on Function 3 by Region

Score / Region	1 N (%)	*2 N (%)	3 N (%)	4 N (%)	5 N (%)	6 N (%)	Total
1	3 (3%)	50 (49%)	37 (36%)	11 (11%)	1 (1%)	0 (0%)	102
2	0 (0%)	24 (25%)	45 (46%)	22 (23%)	5 (5%)	1 (1%)	97
3	13 (12%)	50 (48%)	30 (29%)	8 (8%)	4 (4%)	0 (0%)	105
4	4 (4%)	33 (31%)	53 (50%)	14 (13%)	2 (2%)	0 (0%)	106
5	5 (5%)	24 (26%)	50 (54%)	13 (14%)	1 (1%)	0 (0%)	93
6	3 (3%)	29 (33%)	50 (56%)	6 (7%)	1 (1%)	0 (0%)	89
7	3 (2%)	67 (44%)	68 (45%)	12 (8%)	1 (1%)	0 (0%)	151
8	0 (0%)	17 (18%)	56 (60%)	18 (19%)	2 (2%)	0 (0%)	93
Makeup	0 (0%)	14 (30%)	21 (46%)	9 (20%)	2 (4%)	0 (0%)	46
Total	31 (4%)	308 (35%)	410 (46%)	113 (13%)	19 (2%)	1 (1%)	882

It is interesting to note that, while better than a third of all particpants chose the correct answer, the combined percentages of all participants in any single region who chose acceptable scores ranged from 71% to 91%. Thus, seven in ten participants--at a minimum--selected acceptable scores on this function, but in no region did the percentage of correct responses reach as much as 50%.

Function 4 concerns the teacher's ability to monitor student learning. In some ways it is like Function 2 in that it examines the interaction between teacher and students. Both functions, for example, include practices that require the teacher to monitor students. The difference is that the teacher's motive for monitoring is different. In Function 2 the teacher monitors students' behavior, while in Function 4 the teacher monitors students' learning. The difference is often a subtle one and participants in training experience some difficulty in observing the distinction. Table 7 shows participants' responses on Function 4 of the post-test.

TABLE 7

Distribution of Responses on Function 4 by Region

Score / Region	1 N (%)	2 N (%)	*3 N (%)	4 N (%)	5 N (%)	6 N (%)	Total
1	2 (2%)	67 (64%)	30 (29%)	4 (4%)	1 (1%)	0 (0%)	104
2	1 (1%)	42 (44%)	44 (46%)	8 (8%)	1 (1%)	0 (0%)	96
3	7 (7%)	59 (56%)	37 (35%)	2 (2%)	0 (0%)	1 (1%)	106
4	5 (5%)	47 (44%)	47 (44%)	6 (6%)	1 (1%)	0 (0%)	106
5	6 (6%)	51 (54%)	33 (35%)	4 (4%)	0 (0%)	0 (0%)	94
6	6 (7%)	51 (57%)	31 (35%)	1 (1%)	0 (0%)	0 (0%)	89
7	4 (3%)	100 (66%)	43 (28%)	4 (3%)	0 (0%)	0 (0%)	151
8	2 (2%)	51 (54%)	37 (39%)	4 (4%)	0 (0%)	0 (0%)	94
Makeup	0 (0%)	20 (43%)	20 (43%)	6 (13%)	0 (0%)	0 (0%)	46
Total	33 (49%)	488 (55%)	322 (36%)	39 (4%)	3 (1%)	1 (1%)	886

While 36% of participants chose the correct score point, fully 95% of participants were in the acceptable range, with the preponderance of the 95% rating the teacher's performance on this Function as a 2, as opposed to the normed score of 3.

The fifth function on the TPAI is Instructional Feedback. This function assesses the teacher's ability to inform students about the adequacy or correctness of their in-class and out-of-class work. This function is extremely important in that it also emphasizes how teachers respond to students' answers to in-class questions. If the student, for example, responds incorrectly to the teacher's question, does the teacher offer help to the student, does the teacher re-phrase the question, or does the teacher simply move on to another student? The Teacher Expectation and Student Achievement program, developed by the Los Angeles Unified School District, is one of the few teacher workshop programs that directly deals with this function, and many teachers throughout North Carolina are familiar with it. Not surprisingly, participants in the TPAS workshops were fairly successful in their ability to assess the videotaped teacher's performance on this function, as Table 8 below shows.

TABLE 8

Distribution of Responses on Function 5 by Region

Score / Region	1 N (%)	2 N (%)	*3 N (%)	4 N (%)	5 N (%)	6 N (%)	Tota.
1	1 (1%)	32 (31%)	58 (56%)	12 (12%)	1 (1%)	0 (0%)	104
2	0 (0%)	15 (16%)	52 (54%)	21 (22%)	8 (8%)	0 (0%)	96
3	3 (3%)	26 (25%)	61 (58%)	14 (13%)	1 (1%)	1 (1%)	106
4	3 (3%)	7 (7%)	61 (58%)	29 (28%)	5 (5%)	0 (0%)	105
5	1 (1%)	25 (27%)	56 (60%)	10 (11%)	1 (1%)	0 (0%)	93
6	1 (1%)	17 (19%)	61 (69%)	9 (10%)	1 (1%)	0 (0%)	89
7	5 (3%)	28 (19%)	83 (55%)	29 (19%)	5 (3%)	0 (0%)	150
8	0 (0%)	9 (10%)	68 (73%)	16 (17%)	0 (0%)	0 (0%)	93
Makeup	0 (0%)	5 (11%)	27 (59%)	13 (28%)	1 (2%)	0 (0%)	46
Total	14 (2%)	164 (19%)	527 (60%)	153 (17%)	23 (3%)	1 (1%)	882

On this function, 60% of participants chose the correct answer with an additional 36% choosing answers in the acceptable range. Thus, 96% of participants chose acceptable answers on the post-test on Function 5.

Taking all of these data together, can we make any statements about the capability of participants to use the instrument? Are grounds for improvement indicated?

Two statements seem warranted. First, it seems clear that the vast majority of participants selected the correct, or at least acceptable answers for each function on the post-test. Table 9 presents an extract of data shown in Table 2 that support this statement.

TABLE 9

Summary Results by Function

Function	% Correct	% Acceptable
1	36	88
2	72	98
3	35	85
4	36	96
5	60	96

Second, it would appear from these same data that people who are responsible for observing and evaluating classroom performance of teachers would profit from additional training in aspects of time management, instructional presentation, instructional feedback and, perhaps, instructional monitoring.

In the long run, a larger problem can be foreseen. We have already indicated that, for measurement reasons, a tolerance of ±1 was deemed to be acceptable for establishing ratings. In fact, however, this creates the anomalous situation of one evaluator awarding a 2 to a performance that another observer would say merits a 4. Both would be acceptable, within the logic of our system.

In our present post-test, this tolerance factor resulted in a high degree of success, but a wide range of divergence. Table 10 summarizes the percentages of all participants who awarded scores that are the equivalent of unacceptable or below standard ratings (1 or 2 on the TPAI scale) as compared with those who gave standard or above points (3-6) for the same performance.

TABLE 10

Distribution of Scores on Acceptable/Unacceptable Ratings

(Percentage of Participants Describing Performance as Acceptable/Unacceptable)

Function	Unacceptable	Acceptable	Norm Score
1	37%	63%	(2) Unacceptable
2	80	20	(2) Unacceptable
3	39	61	(2) Unacceptable
4	59	41	(3) Acceptable
5	21	79	(3) Acceptable

This table is provocative precisely because it shows that significant numbers of training participants were unable to agree that the videotape behavior sample was at-standard or above or was below-standard. The norm ratings indicate that the test teacher showed standard performance on only 2 functions: 4 and 5. Yet almost 60% of the participants awarded ratings that were the equivalent of below-standard on Function 4. Conversely, more than 60% of participants felt the sample teacher was at-standard or above on Function 1, while the norm score indicated that performance was below-standard.

It would be ill-advised to react to this difficulty too violently at this time. However, these data do suggest that a two-step evaluation process may be wise. In the first step, the evaluator would determine that performance was at standard or it was not. The second step would involve deciding how much above

or below standard the performance was. Put another way, we would accept a deviation of ±1 only after determining that the teacher was at standard or not. This would yield score ranges of:

1 to 2 or 3 to 4 4 to 5 5 to 6

There may be other solutions to this problem. Nevertheless, we need to acknowledge that the problem exists; that, within our present framework, the problem is not fatal to the system, but that we need to work on a reasonable solution.

Success of Individual Participants

To this point, we have examined the data collected from the 27 workshops using the function as the unit of analysis. A different perspective on the workshops can be rendered by looking at the same data, but using individual performance as the unit of analysis. This somewhat different perspective should illuminate the ability of individuals to utilize the TPAS, or, more precisely, at least enable us to comment on the relative success of the SDPI-sponsored training. Table 11 shows how successful individuals were in terms of acceptable performance on the post-test exercise, completed in the last training session.

TABLE 11

Number and Percentage of Participants Rating All 5 Functions Within Acceptable Range and Outside Acceptable Range, by Regions and by Total

Region	Participants In (Successful)	(%)	Participants Out (Unsuccessful)	(%)	Total	
1	74	(70%)	32	(30%)	106	(12%)
2	61	(62%)	38	(38%)	99	(11%)
3	82	(78%)	23	(22%)	105	(12%)
4	73	(68%)	34	(32%)	107	(12%)
5	69	(72%)	27	(28%)	96	(11%)
6	73	(82%)	16	(18%)	89	(10%)
7	129	(84%)	23	(16%)	152	(17%)
8	63	(67%)	31	(33%)	94	(10%)
Makeup	32	(67%)	16	(33%)	48	(5%)
Total	656	(73%)	240	(27%)	896	(100%)

Of all participants, 73% were successful in assigning acceptable scores in each of the five functions and 27% were not. Within the 27% we include any post-test effort that included incomplete data. That is, if an individual rating sheet did not include 5 ratings (one per function) we counted the data as incomplete and necessarily included that individual in the "Participants Out" column. (Correcting to discount these incomplete papers gives a true "Participants Out" percentage of about 25% of the total.) When we examine the data by regions, we see that the range of success, expressed as a percentage, runs from 62% to 84%, with a median rate of 70%.

These data can also be reported by success among incumbents of three job groups: principals, observer-evaluators, and others. When analyzed thus, the results shown in Table 12, on the next page, are obtained.

At first glance, it appears that observer-evaluators, as a group were more successful than were principals or others. Two mitigating factors should be borne in mind, however. First, the total number of observer-evaluators is quite small, relative to each of the other groups (1:4.75 and 1:6.3). Second, the observer-evaluators were recruited from among the most competent teachers in any given system, whereas the principals and others probably represent a wider range of ability among all incumbents. In any event, the principals as a whole reproduce exactly the success rates of all participants taken together. Since 50% of all participants were principals, we can predict that the success rate of about 75% is generalizable to the whole group of participants, which we know is true.

TABLE 12

Number/Percentage of Participants in Job Roles Within Acceptable and Unacceptable Ranges by Regions

Region	Principal In N (%)	Principal Out N (%)	Observer/Evaluator In N (%)	Observer/Evaluator Out N (%)	Other In N (%)	Other Out N (%)
1	43 (72%)	17 (28%)	1 (100%)	0 (0%)	28 (65%)	15 (35%)
2	17 (55%)	14 (45%)	15 (88%)	2 (12%)	28 (56%)	22 (44%)
3	40 (75%)	13 (25%)	4 (100%)	0 (0%)	34 (77%)	10 (23%)
4	35 (63%)	21 (37%)	5 (100%)	0 (0%)	32 (71%)	13 (29%)
5	36 (78%)	10 (22%)	5 (71%)	2 (29%)	25 (63%)	15 (37%)
6	35 (78%)	10 (22%)	0 (0%)	0 (0%)	36 (86%)	6 (14%)
7	82 (85%)	15 (15%)	11 (85%)	2 (15%)	31 (84%)	6 (16%)
8	35 (67%)	18 (34%)	10 (71%)	4 (29%)	9 (50%)	9 (50%)
Makeup	6 (55%)	5 (45%)	1 (100%)	0 (0%)	24 (69%)	11 (31%)
Total	329 (73%)	123 (27%)	52 (84%)	10 (16%)	247 (70%)	107 (30%)

While a 75% rate of success on a workshop of this type is quite respectable, there is at least one more analysis that will indicate, at least least tangentially, a measure of success. We said earlier that, in order to be considered successful in the training, the participant had to score within the acceptable range on all five function ratings. It would be interesting to know, of those who failed to reach this standard, how any individuals failed on one measure, as opposed to the number failing on two measures or more. In other words, how many participants were successful on 80% of the instrument? Table 13 presents that information by job class by region.

TABLE 13

Number of Participants Out on 1 and Out on 2 or More

	Out On 1				Out On 2 or More				Total			
Region	Prin.	O/E	Other	Total	Prin.	O/E	Other	Total	Prin.	O/E	Other	Total
1	11	0	9	20	6	0	6	12	17	0	15	32
2	7	2	10	19	7	0	12	19	14	2	22	38
3	9	0	6	15	4	0	4	8	13	0	10	23
4	12	0	6	18	9	0	7	16	21	0	13	34
5	8	2	9	19	2	0	6	8	10	2	15	27
6	8	0	4	12	2	0	2	4	10	0	6	16
7	8	2	4	14	7	0	2	9	15	2	6	23
8	13	4	6	23	5	0	3	8	18	4	9	31
Makeup	2	0	5	7	3	0	6	9	5	0	11	16
Total	78	10	59	147	45	0	48	93	123	10	107	240

It is clear that about 60% of those failing to achieve success were, in fact, out of the range on only one function. It will probably be relatively easy to provide remediation for those individuals at the level of the local school system. For the 93 individuals, however, who were out on two or more functions, remediation will probably be more difficult, since it is unclear whether their failure results from inability to adapt to the TPAS, to the TPAI, from an inability to recognize relative quality of the instructional act

or from some reason totally unrelated to the training experience. Performance of these individuals should be carefully evaluated--and monitored--at the local level to avoid inequitable application of the Teacher Performance Appraisal System in their schools or school system.

Conclusion

A number of conclusions seen warranted by the data presented above. It should be remembered, however, that the training was not conducted as a research project. The training was presented as a service to local education agencies. There was no hypothesis being tested in the training; there were no experimental nor control groups; there was no attempt made to establish laboratory conditions; there were either dependent nor independent variables being studied. Our conclusions, then, are the result of logically examining the data available. The discussion that follows each conclusion will help the reader understand that logical examination.

Conclusion 1: The training experience was successful.

By and large, this is the primary conclusion one can draw from the data. whether we analyze these data from the perspective of the functions (Table 4-9) or from the perspective of the participant (Table 11-13), it is clear that most participants (73%) in the training, learned enough in the four days of the training to make five separate decisions--one per function--within an acceptable range. In addition to these participants, an additional 15% of participants were able to select acceptable scores on four out of five of the function decisions. Only about 10% of participants rated outside the acceptable range on two functions or more.

Conclusion 2: Success rates would have been higher if consensus decision, rather than individual ratings, had been used.

This conclusion springs from an examination of the experiences reported by trainers. During the process of training participants to make ratings, a consensus process was used as an intermediate step toward establishing rater independence. When viewing video-taped teaching episodes, participants were asked to rate the tape and then to compare and discuss their ratings with other participants in their training groups in order to arrive at a consensus-based rating. Almost invariably, the process of consensus-building helped individuals understand their own ratings better, helped them recall significant features of the teaching episode that they might have overlooked, and generally resulted in agreements within consensus groups that were more accurate than had been some individual ratings.

Of course, in the world of the schools, the principal is legally responsible for the observation/evaluation of teachers' performance. If this responsibility were shared among several trained observers/evaluators, including the principal, all of whom used the same evaluative criteria, then we would expect to see the ratings of performance become increasingly precise. Indeed, among some of the Career Ladder Development pilot units variations of this consensus approach to evaluation are evolving.

An additional benefit of a team-approach to evaluation is that it spreads the work out over more evaluators. One of the most commonly expressed concerns of principals during the training was the amount of time required for observation and evaluation. This concern is very real. If, for example, a pre-observation conference requires 30 minutes, an observation requires 60 minutes, a post-observation conference requires 45 minutes, with an additional

45 minutes for the required data analysis, and the actual evaluation requires 60 minutes, the following formula will result:

Observation #1 (Announced):	180
Observation #2 (Unannounced):	150
Observation #3 (Announced):	180
Evaluation:	60
	570 minutes

If a school has 50 teachers and only one administrator who supervises teacher personnel, then about 60 days of 8 hours will be required to complete all observations and evaluations. This "worst-case" scenario is exacerbated by school calendars that require these 60 days to fall with a 150-day time frame required by the fact that observations early and late in the year will be impractical. The answer to the problem does not lie in abandoning the system of observations and evaluation based on performance. Rather, the answer lies in providing additional help--either in the form of system-supported observers/evaluators, assistant principals, or other building-level administrators--in carrying out this task.

Conclusion 3: The wide range of acceptable scores should be carefully considered and perhaps, revised.

We have illustrated the problem above (pages 4-5). Essentially, it is this: we must use a rating scale that allows some measure of tolerance in rating selection. By accepting ratings of ±1 in training, we made ranges whose discrete score points were functionally identical. That is, a score range of 1-3, 2-4, 3-5, or 4-6, for example, represent equally acceptable ratings. Clearly, ratings of 2 and 4 or 3 and 5 or 1 and 3 are not equivalent.

The answer to this problem may be to use a two-stage decision for assigning ratings. In the first stage, the rater must decide whether performance on the function is acceptable or unacceptable. Then, the rater must decide the degree to which performance is acceptable or unacceptable. There would be no acceptable tolerance on the first step, a tolerance of ±1 could be admitted in the second stage.

Participants in this training were not asked to use this two-stage process. However, a reexamination of the data tabulated in Table 10 shows that on two functions (#2 and #5) the large majority of participants gave ratings on the appropriate side of the acceptable/unacceptable scale. However, on three functions, the majority decided on the inappropriate side of the scale. However, the group differences were much smaller on these 3 functions. This suggests that, if the training had emphasized a two-stage process, we probably could have reduced discrepancies even more than was true. This process would minimally have the effect of clarifying the differences between acceptable and unacceptable performance.

Conclusion 4: Participants' prior training and experiences probably influenced TPAS training outcomes.

This conclusion is not surprising, but its implications should not be overlooked. No real attempt was made to determine participants' prior experiences or training. However, we know that participants are likely to have participated in any of a number of training experiences that have been promoted throughout North Carolina in recent years. If we think about a few of these major training efforts--Classroom Organization for Effective Teaching, Teacher Expectations and Student Achievement, Assertive Discipline, and Hunter's Instructional Presentation--we notice that the first three place heavy emphasis on student behavior. Moreover, public expectations have placed

emphasis on the same area of school management. Not surprisingly, the function on which participants demonstrated the highest absolute and relative success was Function 2: Management of Student Behavior. A total of 72% of participants gave the exact rating on this function and 98% (an additional 26%) were within the acceptable range.

It would appear, then, that participants' expertise--a combination of training, experience, and the perception of role expectations, perhaps--leads to increased accuracy in evaluation of the teacher's ability to perform a function. Clearly, increased attention through inservice training, clarification of school district and public expectation, can lead to increased skill in recognizing, evaluating, and promoting the practices and functions associated with school effectiveness.